Automated Research Paper Categorization

Introduction

Automated research paper categorization involves the process of classifying research papers into predefined categories based on their titles and abstracts. This task is a critical component of many academic and industrial applications where efficient and accurate categorization of large volumes of documents is required.

Problem Statement

The objective is to classify research papers into one or more of the 57 given categories based on their title and abstract. This is a multi-label text classification problem with the Macro F1 score as the judging criteria.

Data Description

The training dataset consists of 51,210 samples with the following columns:

- **Id**: Unique identifier for each paper
- **Title**: Title of the research paper
- **Abstract**: Summary of the research paper
- **Categories**: List of categories to which the paper belongs

Dataset: Kriti 2024 | Kaggle

Data Cleaning

1. **Parsing Categories**: Converted the "Categories" column from a string to a list data type.
2. **Creating Text Column**: Combined the "Title" and "Abstract" columns into a single "Text" column.
3. **Text Cleaning**:
   - Parsed LaTeX to text using the `pylatexenc` library.
   - Removed redundant spaces and punctuations.
   - Converted the text to lowercase.

Data Analysis

The data analysis revealed that the categories exhibit a long-tailed distribution, indicating a class imbalance. This insight is crucial for developing an effective model.

Approaches

To tackle the problem, several state-of-the-art transformer models were employed:

1. **BERT-based Transformers**:
   - **bert-base-uncased**
   - **scibert-scivocab-uncased**
2. **T5 Transformer**:
   - A transformer model designed for text-to-text tasks, used for its flexibility in handling various NLP tasks.
3. **Class Imbalance Solution**:
   - SciBERT with a special loss function tailored for long-tailed distributions to address the class imbalance issue.

Model Training and Results

The performance of the models was evaluated based on the Macro F1 score. The results for different models and configurations are as follows:

1. **bert-base-uncased**:
   - Trained for 2 epochs, achieved a Macro F1 score of 0.49.
2. **scibert-scivocab-uncased**:
   - Trained for 6 epochs, achieved a Macro F1 score of 0.65.
   - With a special loss function for class imbalance, it also achieved a Macro F1 score of 0.65.
3. **T5 Transformer**:
   - Trained for 2 epochs, achieved a Macro F1 score of 0.60.

Google Colab

https://colab.research.google.com/drive/1xYIm0ASg74oLUxIFa6pqPlu37×8hrZ2z?usp=sharing

Final Output :

Output.csv

Conclusion

This report presents a detailed analysis and implementation strategy for automated research paper categorization using advanced transformer models. The results demonstrate the effectiveness of different models in handling the multi-label classification problem, with special attention given to addressing class imbalance.

For further improvement, more sophisticated models and techniques for handling class imbalance can be explored. Continuous fine-tuning and validation against diverse datasets will enhance the robustness and accuracy of the categorization system.